Causal discovery for dopamine transporter haplotype and reward - related brain activation for adult ADHD

Elena Sokolova[1], Perry Groot[1], Tom Claassen[1], Daniel von Rhein[2], Jan Buitelaar[2] and Tom Heskes[1]

[1]Faculty of Science, Radboud University Nijmegen, The Netherlands
[2]Donders Institute for Brain, Cognition and Behaviour, Radboud University Medical Center Nijmegen, The Netherlands
* Corresponding author, e-mail: e.sokolova@cs.ru.nl

1. Introduction

Existing algorithms in causal discovery deal reasonably well with models that contain only discrete variables or only Gaussian variables, while real-world data often contains mixture variables, where continuous variables are not Gaussian.

2. Proposed method

Step 1 Mixture of discrete and continuous variables

For each variable X_i estimate the rescaled empirical distribution

$$\hat{F}_i(x) = \frac{1}{n+1}\sum_{i=1}^{n}\mathcal{I}\{X_i < x\}, \qquad (1)$$

and then transform the data into Gaussian normal scores $\hat{X}_i = \hat{\Phi}_i^{-1}(\hat{F}(X_i))$. In this step missing values are ignored.

Step 2 Correlation matrix with missing data

Use Pearson correlation to estimate dependencies between variables, since data is now Gaussian. Use the expectation maximization algorithm to estimate the correlation matrix with missing data.

Step 3 Apply BCCD

Estimate the causal structure of the graph using BCCD [1]. Estimate the reliability of the causal relations, using the Bayesian Information Criterion (BIC):

$$BICscore(\mathbf{D}|\mathcal{G}) = M\sum_{i=1}^{n} -\frac{1}{2}\log\frac{|\Sigma|}{|\Sigma_{Pa_i}|} - \frac{\log M}{2}\mathrm{Dim}[\mathcal{G}] \ , \qquad (2)$$

where n is the number of variables, M is the sample size, $\mathrm{Dim}[\mathcal{G}]$ is the number of parameters in the model corresponding to graph $\mathcal{G}$, Pa_i are the parents of node X_i, and Σ_{Pa_i} is a correlation matrix between the parents of variable X_i.

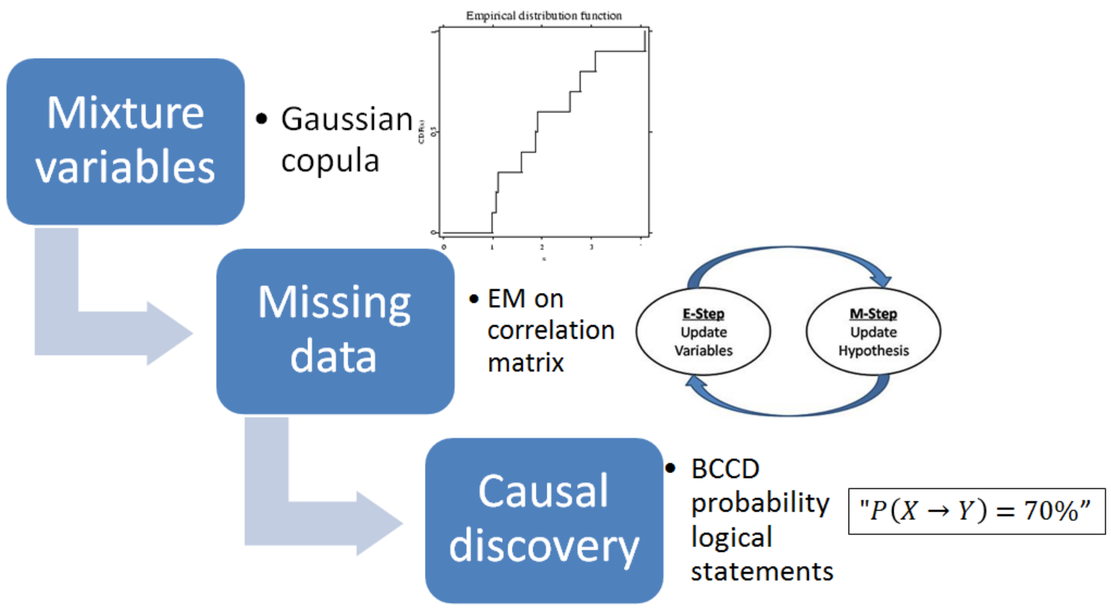

Figure 1: *Three steps of the algorithm for causal discovery with missing data and mixture variables*

3. ADHD results

We have applied the BCCD algorithm with EM to the ADHD data set that was collected as a part of the NeuroIMAGE study [2]. The goal of our study is to identify the endophenotypic model that explains the relationships between genes, brain functioning, behaviors, and disease symptoms. A causal network learned from the data is presented in Figure 2.

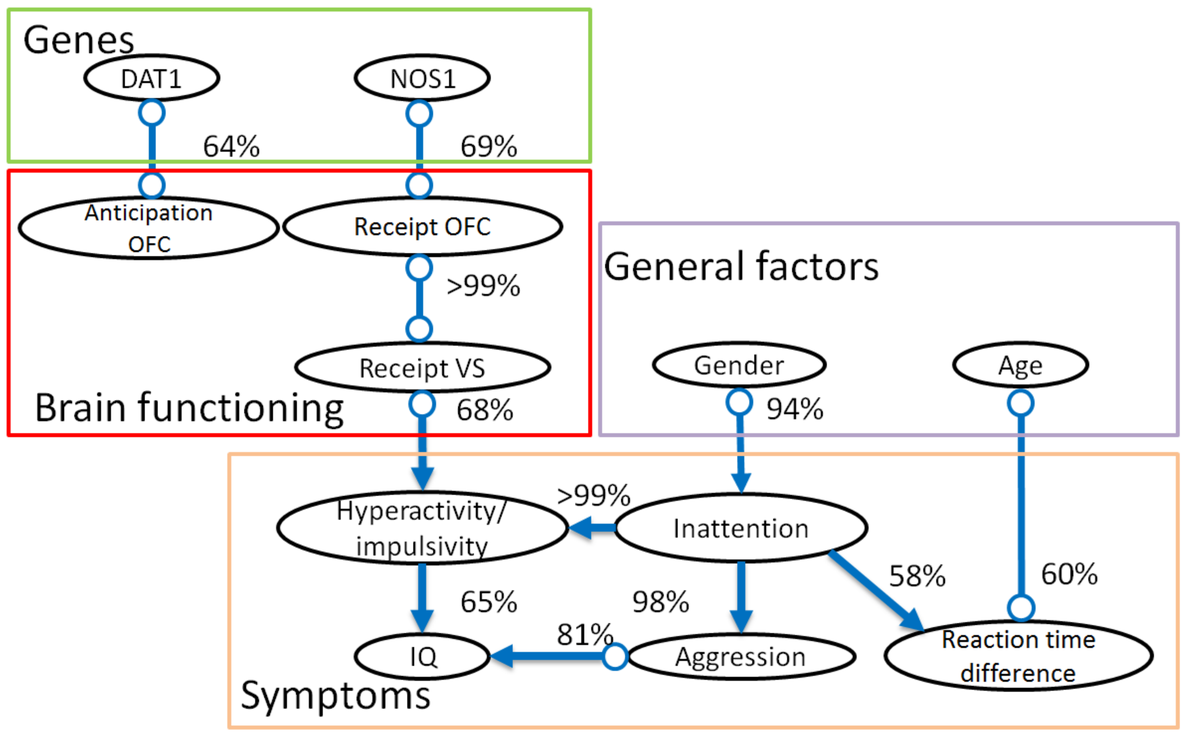

Figure 2: *The causal graph representing causal relationships between variables for the ADHD data set. The graph represents a PAG, where edge directions are marked with " − " and " > " for invariant edge directions and with "○" for non-invariant edge directions. The reliability of an edge between two variables is depicted with a percentage value near each edge.*

The graph built by BCCD shows an effect of genes on brain functioning, the effect of brain functioning and general factors on disease symptoms, and an interaction between these symptoms. This model confirms several causal paths that were previously presented in other studies, but also suggests new endophenotypic pathways.

References

[1] T. Claassen and T. Heskes. A Bayesian approach to constraint based causal inference. In *Proceedings of the UAI Conference*, pages 207–216, 2012.

[2] D. von Rhein, M. Mennes, H. van Ewijk, A. P. Groenman, M. P. Zwiers, J. Oosterlaan, D. Heslenfeld, B. Franke, P. J. Hoekstra, S. V. Faraone, et al. The NeuroIMAGE study: a prospective phenotypic, cognitive, genetic and MRI study in children with attention-deficit/hyperactivity disorder. Design and descriptives. *European child & adolescent psychiatry*, pages 1–17, 2014.

Acknowledgement: The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grand agreement n° 278948.